Principal Component Analysis for Interval Data

Paula Brito

Fac. Economia & LIAAD-INESC TEC, Universidade do Porto

ECI 2015 - Buenos Aires
T3: Symbolic Data Analysis:
Taking Variability in Data into Account

Outline

Outline

Principal Component Analysis: Objectives

Principal Component Analysis applies to data described by (p) numerical variables.

Objective:

to obtain an approximate representation of the individuals in a sub-space of dimension $q < p$.

That is, to represent them with **a small set of variables**

The representation in a space of lower dimension is done by a projection - and NOT by variable selection.

$\longrightarrow$ **Dimension reduction**

Principal Component Analysis: Objectives

The choice of the new representation space is done according to the following criterion :

The average squared distance between projected points (measuring their dispersion) must be as large as possible.

We wish to distort the point configuration as little as possible and therefore also the distances between them (they can only decrease).

In other words:

the dispersion of the projected point set should be maximized

Principal Component Analysis: Objectives

From the variables point of view,
the problem consists in determining **new variables**
linear combinations of the original variables (previously centered)
non-correlated among themselves, and with maximum variance

The solution is nested :

The best subset of q (new) variables is obtained from the best subset of $q-1$ variables,

together with the best aditional variable, non-correlated with them, that maximizes the dispersion in the projection.

Principal Component Analysis: Solution

The solution of the problem is given by the (q) normed eigenvectors of the variance-covariance matrix V, in non-normed PCA

or of the correlation matrix R, in normed PCA

associated to the (q) largest eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$$

The total dispersion of the projected dataset

$=$ sum of the variances of the new variables is given by:

$$\sum_{\alpha=1}^{q} \lambda_\alpha$$

Principal Component Analysis: Properties

- The principal components are variables with null mean
- The principal components are non-correlated variables
- The variance da α_{th} principal component c_α is equal to the correponding eigenvalue
- The principal components are linear combinations of the centered (or standardized) original variables, defined by the eigenvectors associated with the corresponding eigenvalues

If an important part of the dispersion is explained by a small number q of principal components:

we may use just some of them for interpretation and future analysis, instead of the original p

Outline

Principal Component Analysis of Interval data

- Factorial method for interval data
- Graphical representation by rectangles in the factorial planes
- Interpetation indices

Principal Component Analysis of Interval data

Direct solution:

- use covariance / correlation values obtained for the interval-valued variables
- proceed as for real (classical) data
- obtain interval representations in the new variables' space - applying the linear combinations to lower and upper bounds of the observed data

Principal Component Analysis of Interval data

Two first methods

(Cazes, P., Chouakria, A., Diday, E., Schektman, Y. (1997)) :

- Vertices method:
  Analysis of the data array with $n \times 2p$ rows and p coloumns containing all vertices of the hyper-rectangles in the interval data-array
- Centres method:
  Replaces each interval by its centre

See also: Chouakria, A., Billard, L., Diday, E. (2011)

SPCA: Vertices method

Example:
if s_i has the description $([a, b], [c, d])$, s_i will be "transformed" in :

$$\begin{bmatrix} a & c \\ b & c \\ a & d \\ b & d \end{bmatrix}$$

A classical PCA of the new $(n \times 2p)$ rows $\times p$ columns data array is performed.

Weight of $s_i = \frac{p_i}{2^p}$ (p_i : original weight of s_i)

SPCA: Vertices method

Let Y_j^* be the les q first principal components, $j = 1, \ldots, q$

For each principal component Y_j^*:

the minimum value $min_j(s_i)$ and the maximum value $max_j(s_i)$ of Y_j^* are detemined, accross the $2p$ rows representing each entity s_i

We then define: $Y_j^*(s_i) = [min_j(s_i), max_j(s_i)]$

SPCA: Vertices method

Interpretation indices: Quality of representation

G : centre of gravity of the $n \times 2p$ points
d : Euclidean distance
L_i : $2p$ points representing s_i

Contribution of the $2p$ vertices representing s_i to the total sum of squares:

$$COR^1(i, v_j) = \frac{\sum_{k \in L_i} Y_{kj}^{*2}}{\sum_{k \in L_i} d^2(k, G)}$$

Mean of the squared cossinus of the angles between the $2p$ vertices representing s_i and the factorial axis v_j:

$$COR^2(i, v_j) = \frac{1}{2^p} \frac{\sum_{k \in L_i} Y_{kj}^{*2}}{\sum_{k \in L_i} d^2(k, G)}$$

SPCA: Vertices method

Interpretation indices: Contributions

Contribution of s_i to the variance of the j^{th} principal component:

$$CTR(i, v_j) = \frac{\sum_{k \in L_i} q_k Y_{kj}^{*2}}{\lambda_j} = \frac{p_i}{2^p \lambda_j} \sum_{k \in L_i} Y_{kj}^{*2}$$

Contribution of s_i to the total inertia:

$$INR(i) = \frac{p_i}{2^p} \frac{\sum_{k \in L_i} d^2(k, G)}{\sum_j \lambda_j}$$

SPCA: Centres method

Only the intervals' midpoints are used.

Example:
if s_i has the description $([a, b], [c, d])$, s_i will be "transformed" in :

$$c_i = \left(\frac{a+b}{2}, \frac{c+d}{2} \right)$$

A classical PCA of the new $n \times p$ data array is performed

x_{ij}^c : midpoint of variable Y_j for entity s_i (centre c_i)

$\overline{x_j^c}$: mean of the midpoints of variable Y_j

$v_\ell = (v_{1\ell}, \ldots, v_{p\ell})$: the ℓ^{th} eigenvector

The ℓ^{th} principal component of centre c_i is given by:

$$Y_{i\ell}^{*c} = \sum_{j=1}^{p} (x_{ij}^c - \overline{x_j^c}) v_{j\ell}$$

SPCA: Centres method

I_{ij} : interval corresponding to entity s_i and variable Y_j

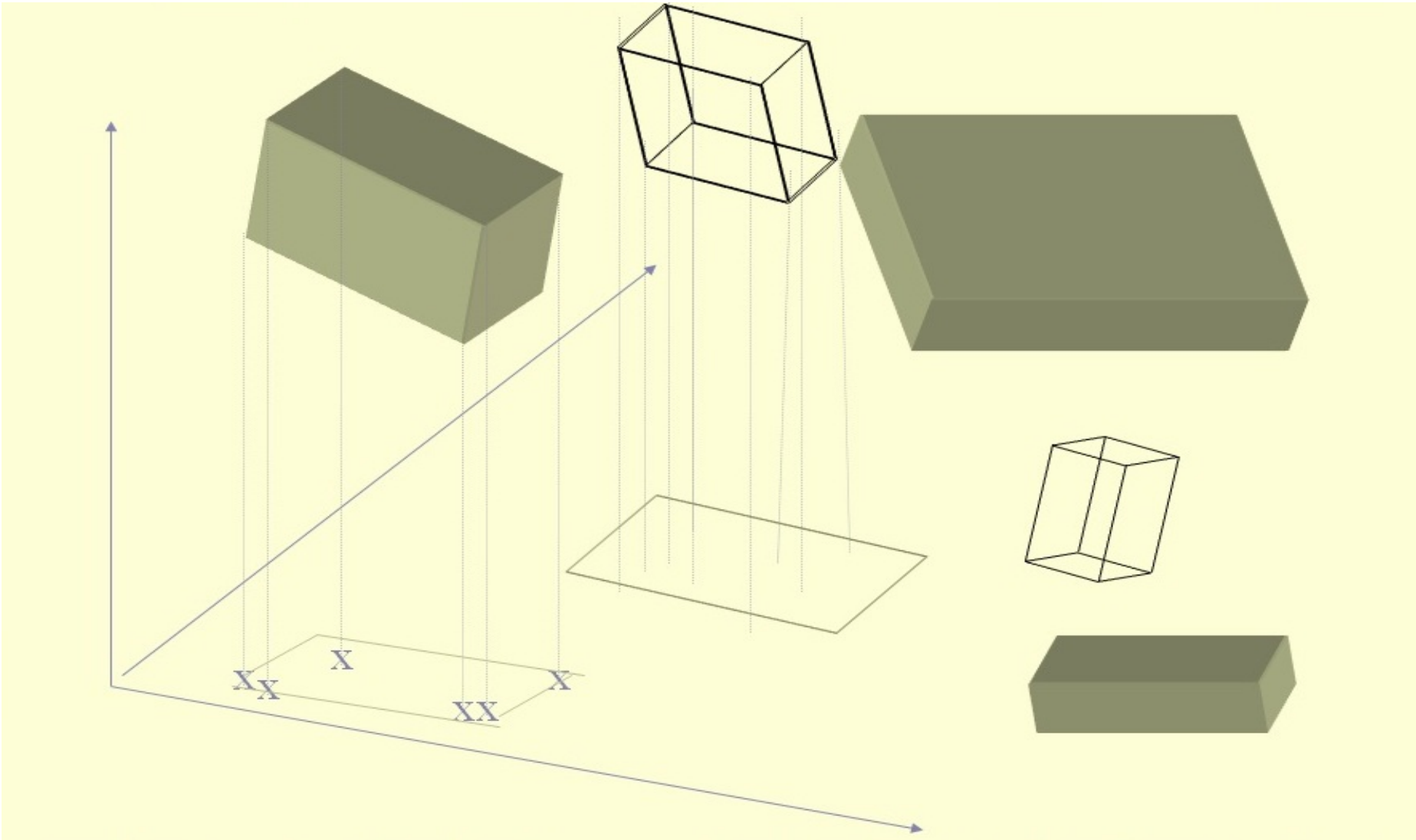
Interval corresponding to s_i for the ℓ^{th} principal component:

$$I_{ij}^* = [\underline{I_{ij}^*}, \overline{I_{ij}^*}]$$

$$\underline{I_{ij}^*} = \sum_{j=1}^{p} \text{Min}_{x \in I ij}\{(x - \overline{x_j^c})v_{j\ell}\}$$

$$\overline{I_{ij}^*} = \sum_{j=1}^{p} \text{Max}_{x \in I ij}\{(x - \overline{x_j^c})v_{j\ell}\}$$

Principal Component Analysis of Interval data



Principal Component Analysis of Interval data

Other methods for PCA of Interval data:

- Lauro, C., Palumbo, F. (2000). Principal component analysis of interval data:a symbolic data analysis approach.
- Palumbo, F., Lauro, C. (2003). A PCA for interval valued data based on midpoints and radii.
- Gioia, F., Lauro, C. (2006). Principal component analysis on interval data.
- Irpino, A. (2006). "Spaghetti" PCA analysis: An extension of principal component analysis to time dependente interval data.